Women and Ischemia Syndrome Evaluation (WISE)
Diagnosis and Pathophysiology of Ischemic Heart Disease
Workshop

October 2-4, 2002

Session I

1. Topic and Author

Epidemiologic Issues in the Clinical Diagnosis of Angina. George A. Diamond, MD, FACC

2. Where we stand in 2002. Overview/rationale for inclusion of topic.

Cardiac diagnostic testing is becoming progressively more complex as the number of procedures available to the physician—and the uses to which they are put—continue to increase. One consequence of this technologic explosion is that test interpretation is made more difficult by the frequent occurrence of discordant results. At such times, rational judgments can be compromised if words alone are used to describe complex beliefs. For example, when 205 subjects were asked to assign a numeric probability to the meaning of the word "often", their estimates ranged from a low of only 0.2 to a high of 0.9 (1-4). If the meaning of such words is so variable, how shall we best ensure the accuracy of our judgments?

Bayes' theorem is the formal rule by which one integrates the interpretation of any combination of observations in light of past experience (5). I outline herein the conceptual importance of Bayes' theorem to clinical test interpretation, and show how it can be used to help the physician interpret tests for the diagnosis and evaluation of coronary artery disease.

Bayes Theorem. The conventional measures of test accuracy are called sensitivity and specificity. Sensitivity (also called true positive rate) measures a test's ability to correctly indicate the presence of disease. Numerically, it is the frequency of a positive test result in a population with disease. Specificity (also called true negative rate) measures a test's ability to correctly indicate the absence of disease. Numerically, it is the frequency of a negative test result in a population without disease. These definitions, therefore, separate a tested population into four subsets—two test result subsets ("positive" and "negative") and two diagnostic subsets ("disease" and "nondisease"). These subsets are illustrated in table 1.

Although sensitivity and specificity define a test's inherent accuracy, its ultimate interpretation depends on a third variable—the prevalence of disease in the tested population. Numerically, prevalence is the frequency of disease in the population. For example, consider a population of 100 patients with an intermediate disease prevalence of 50%—50 patients with disease and 50 patients without disease. If we evaluate each of these patients with a test that has a 70% sensitivity and 90% specificity we would expect the following:

50 x 0.7 = 35 True Positive Test Responses

50 x (1-0.9) = 5 False Positive Test Responses

There are, therefore, a total of 40 positive test responses, only 35 of which occurred in diseased patients. The prevalence of disease in the population of patients with a positive test response is therefore 35/(35+5) or 88%. Similarly, the probability of disease for any given patient with a positive test is also 35/(35+5) or 88%. The prevalence of disease in a population, then, is operationally equivalent to the probability of disease in any individual member of that population.

Likewise, if we analyze the population of negative test responders, we would expect the following:

50 x (1-0.7) = 15 False Negative Test Responses

50 x 0.9 = 45 True Negative Test Responses

There are a total of 60 negative test responses, and 15 of these occurred in patients with disease. The prevalence of disease in the population of patients with negative test responses is 15/(15+45) or 25%, and the probability of disease for a patient with a negative test response is also 15/(15+45) or 25%. Again, prevalence is equivalent to probability.

These probabilistic outcomes for a positive test response (P+) and a negative test response (P-) can be calculated directly using a simple formula based on Bayes' theorem of conditional probability:

$$P+ = \frac{\text{Sensitivity} \times \text{Prevalence}}{\text{Sensitivity} \times \text{Prevalence} + (1\text{-Specificity}) \times (1\text{-Prevalence})}$$

$$P- = \frac{(1\text{-Sensitivity}) \times \text{Prevalence}}{(1\text{-Sensitivity}) \times \text{Prevalence} + \text{Specificity} \times (1\text{-Prevalence})}$$

Our example illustrates two important features of all diagnostic tests. First, a positive (or abnormal) test does not establish the presence of disease; it only increases its probability. Second, a negative (or normal) test does not exclude the presence of disease; it only lessens its probability. Only if a diagnostic test were perfect—and none is—can the test result be accepted without question. Table 2 summarizes the probability of disease given a positive or negative test result for a range of disease prevalence prior to testing. Thus, if the sensitivity and specificity are known, Bayes' theorem provides a probabilistic interpretation of any test observation as a function of the probability of disease before the test is performed. Note that when prior disease probability is very high (*e.g.*, over 90%) or very low (*e.g.*, under 10%), the test is of limited value. All diagnostic tests are of most value when disease probability is intermediate (e.g. around 50%)—when we are most uncertain. In a high prevalence population, a positive test response serves to confirm the presence of disease, while a negative response does not exclude disease. Likewise, in a low prevalence population, a negative test serves to confirm the absence of disease, while a positive test does not establish disease presence.

Estimating CAD Probability. The probability of coronary artery disease can be estimated from the patient's age, sex, and symptom classification. One widely used classification system is based on three readily determined historical characteristics that are generally accepted as being typical of ischemic cardiac discomfort:

Is the discomfort substernal?
Is it precipitated by exertion?
Is there prompt relief by rest or nitroglycerin?

When all three of these questions are judged by the physician to have been answered in the affirmative, the patient's discomfort is interpreted as typical angina. When only two of the three answers are affirmative, the discomfort is interpreted as atypical angina. When fewer than two answers are affirmative, the discomfort is interpreted as nonanginal. Table 3 summarizes the probability of coronary artery disease based upon a broad review of the medical literature (5). This model has been validated in a number of investigations (6-22).

Verification Bias. This classification schema was developed in a population of approximately 5,000 patients undergoing coronary angiography because of suspected coronary artery disease in the decade between 1966 and 1976 (prior to the widespread use of nuclear stress testing, myocardial revascularization and preventive agents such as ACE inhibitors, beta blockers, aspirin and statins). Less than one third of these patients were reported to have undergone electrocardiographic stress testing. It is likely then that these patients were selected for diagnostic verification in ways very different from those currently used (in the WISE population, for example).

In this context, estimates of test accuracy are often highly distorted by the differential referral of positive and negative test responders for diagnostic or prognostic verification (23-25)—an affirmative consequence of the exercise of good clinical judgment. Diagnostic tests for coronary artery disease, for example, are usually verified by referral for coronary angiography. But only a small fraction of patients suspected of having coronary artery disease are actually referred for angiography, and those who are referred often are not typical of the larger population that they come to represent (26). Thus, patients who are selected for angiography on clinical grounds tend to have more abnormal clinical findings and more extreme test responses than those not so selected—whether or not they have disease. This bias causes a systematic overestimation of diagnostic sensitivity, and an underestimation of diagnostic specificity (24,26).

Suppose you have a diagnostic test with a sensitivity of 70% and a specificity of 90%. Suppose further that you so rely on this test, that you refer each and every patient with a positive test response for diagnostic verification, but you never

refer a patient with a negative test response for verification. Because only positive test responders will undergo verification, every diseased patient will have a positive test (observed sensitivity=100%), but so will every non-diseased patient (observed specificity=0%).

Now suppose that this same test has a prognostic sensitivity of 70% and a prognostic specificity of 90% over a specific duration of follow-up. Suppose further that you refer each and every patient with a positive test response for treatment (and away from prognostic verification through longitudinal follow-up), and that you never refer a patient with a negative test response for treatment. Because only negative test responders will undergo prognostic verification, every patient who does not manifest a clinical event during the follow-up period will have a negative test (observed specificity=100%), but so will every patient who does manifest an event (observed sensitivity=0%).

Thus, whenever the proportion of patients with a positive test response who are referred for verification is different from the proportion of patients with a negative test response, the observed sensitivity and specificity are different from the actual sensitivity and specificity (27-29). This so called verification bias (variably called selection bias, post-test referral bias, and work-up bias) produces directionally opposite effects on sensitivity and specificity with respect to diagnosis and prognosis.

Verification bias can have a major effect on observations in the WISE. Suppose physicians are predisposed to believe that women often have highly atypical symptoms suggestive of ischemic heart disease. As a result, they might be inclined to refer any women with "squirrelly" symptoms for an exercise SPECT study, and to refer any of those with a positive study for coronary angiography. Every woman who is thereby documented to have coronary artery disease will also have these "squirrelly" symptoms, even though the published data on the frequency of nonanginal chest discomfort predicts a very low frequency (table 3). This does not justify treating women with such highly atypical symptoms as being at risk for coronary artery disease.

Diamond et al. have developed a strategy for quantifying the sensitivity and specificity of a test (24) using the probability of disease derived from age, sex, symptom classification, and the results of previous noninvasive testing as a surrogate for angiographic verification (24). Similarly, Begg and Greenes have described a method to correct estimates of sensitivity and specificity that are distorted by this bias, assuming that verification is not conditioned on diagnostic or prognostic outcome independent of the test result, and that predictive accuracy of the test is thereby invariant with respect to verification bias (30). This method explains the observed variability in sensitivity and specificity of exercise electrocardiography among patients undergoing coronary angiography by the preferential referral of abnormal test responders for diagnostic verification (27), and provides a suitable method to correct biased estimates of sensitivity and specificity (31-36).

In summary:

(i) Preferential referral of positive or negative test responders for diagnostic verification can seriously distort (bias) empirical estimates of test sensitivity and test specificity (25);

(ii) these distortions can be mitigated in various ways (24,30,32,37) by considering the distribution of test responses in the unverified patient cohort (debiasing);

(iii) additional consideration of ancillary clinical observations (covariates) can improve the accuracy of these debiased estimates (30-32,38,), but the magnitude of this improvement is not necessarily statistically significant or clinically important (31-32);

(iv) receiver-operating characteristic (ROC) curve area, regardless of the particular method of its determination, is comparatively insensitive to verification bias (39-40).

3. Current challenges and the most important issues for future research

The key assumption underlying the historical evaluation of patients for symptoms of myocardial ischemia is that earlier and more accurate diagnosis of the underlying obstructive coronary artery disease will lead to more appropriate utilization of tests and treatments, thereby resulting in better clinical and economic outcomes. Two factors cast doubt on this reasoning:

- Clinical diagnostic tests (including symptom classification schemas) developed in angiographic populations cannot be applied to non-angiographic populations without adjusting for the distorting effects of verification bias.

- Even if a valid symptom classification schema—applicable to patients *prior* to the decision to refer for stress testing or coronary angiography—were developed, it is now well-recognized that symptoms are a very late manifestation of atherosclerotic disease, and that coronary events often develop as a consequence of the destabilization of a hemodynamically insignificant atherosclerotic plaque. Such plaques are clinically silent. They do not cause

symptoms and cannot be reliably detected by even the most sophisticated noninvasive exercise tests. Thus, even if accuracy is no longer an issue, the clinical relevance of this effort is still open to question.

What is most needed then is an accurate and clinically relevant approach to the triage of patients for assignment to prospectively validated optimal age and sex specific management strategies for the prevention of ischemic events, maintenance of quality-of-life, and maximization of cost-effectiveness.

4. Current challenges in the areas of communicating messages to health care community, patients and the public

Published reports often emphasize statistical significance (p-values) over clinical importance (magnitude of benefit). At the same time, the current trend toward larger and larger clinical trials has unearthed a number of limitations in the conventional assignment of statistical significance (41,42). Thus, because these so-called "megatrials" are often cited as the authoritative foundation for evidence-based practice policies, their underlying credibility is open to question and deserving of a critical reappraisal. Toward this end, we might enlist federal agencies such as the National Institutes of Health, Food and Drug Administration, Health Care Financing Administration, Department of Veterans Affairs, and Institute of Medicine to empanel a task force—along the lines of the Consolidated Standards of Reporting Trials (CONSORT) group (43)—comprising clinical trialists, health outcomes researchers, epidemiologists, statisticians, journal editors, and policy makers. The task force would be mandated to define the theoretical and practical standards for the conduct and reporting of clinical trials (supported, perhaps, by scientific comparisons of previously published empirical data and by reasonable computer simulations). In the course of doing so, the task force would standardize representations of prior probability, and integrate the observed magnitude of treatment effect (absolute and relative risk reductions) with this background information. Appropriately vetted statistical software instantiating these standards could be developed and disseminated via the Internet.

5. Translating new findings to improved diagnosis and treatment/saving lives.

Currently, there is a major disconnect among the various partisan sectors involved in health care (patients, payers, providers) regarding what we *are* doing (the *descriptive* perspective) versus what we *should be* doing (the *prescriptive* perspective). The former is guided more by financial incentive (reimbursement being greater for procedures than for preventive care) and by the conventional mediolegal "standard of care" (behavioral norms) than by the scientific "standard of care" (clinical trial evidence). The greatest challenge for the future will be to find ways to overcome this disconnect, and to develop politically acceptable, clinically realistic incentives to encourage optimal evidence-based management strategies (44-47).

6. References.

1. Robertson WO. Quantifying the meaning of words. JAMA 249:2631-2632, 1983.
2. Bryant GD, Norman GR. Expressions of probability: words and numbers. New Engl J Med 302:411, 1980.
3. Kenney RM. Between never and always. New Engl J Med 305:1098-1099, 1981.
4. Toogood JH. What do we mean by usually? Lancet 1:1094, 1980.
5. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. New Engl J Med 300:1350-1358, 1979.
6. Weintraub WS, Maderia SW, Bodenheimer MM, Seelaus PA, Katz RI, Feldman MS, Agarwal JB, Banka VS, Helfant RH. Critical analysis of the application of Bayes' theorem to sequential testing in the noninvasive diagnosis of coronary artery disease. Am J Cardiol 54:43-49, 1984.
7. Diamond GA, Staniloff HM, Forrester JS, Pollock BH, Swan HJC. Computer-assisted diagnosis in the noninvasive evaluation of patients with suspected coronary artery disease. J Am Coll Cardiol 1:444-455, 1983.
8. Staniloff HM, Diamond GA, Pollock BH. Probabilistic diagnosis and prognosis of coronary artery disease. J Cardiac Rehab 4:518-529, 1984.
9. Wong DF, Tibbits P, O'Donnell J, Collison H, LaFrance N, Han S, Otto A, Karam M, Camargo EE, Wagner HN. Computer-assisted Bayesian analysis in the diagnosis of coronary artery disease. J Nucl Med 23:P83, 1982.
10. Greenberg PS, Ellestad MH, Clover RC. Comparison of the multivariate analysis and CADENZA systems for

determination of the probability of coronary artery disease. Am J Cardiol 53:493-496, 1984.
11. Detrano R, Leatherman J, Salcedo EE, Yiannikas J, Williams G. Bayesian analysis versus discriminant function analysis: their relative utility in the diagnosis of coronary artery disease. Circulation 73:970-977, 1986.
12. Hlatky M, Botvinick E, Brundage B. Diagnostic accuracy of cardiologists compared with probability using Bayes' rule. Am J Cardiol 49:1927-1931, 1982.
13. Melin JA, Wijns W, Vanbutsele RJ, Robert A, DeCoster P, Brasseur LA, Beckers C, Detry JMR. Alternative diagnostic strategies for coronary artery disease in women: demonstration of the usefulness and efficiency of probability analysis. Circulation 71:535-542, 1985.
14. Detry JMR, Robert A, Luwaert RJ, Rousseau MF, Brasseur LA, Melin JA, Brohet CR. Diagnostic value of computerized exercise testing in men without previous myocardial infarction. A multivariate, compartmental and probabilistic approach. Eur Heart J 6:227-238, 1985.
15. Diamond GA, Forrester JS. Improved interpretation of a continuous variable in diagnostic testing: probabilistic analysis of scintigraphic rest and exercise left ventricular ejection fractions for coronary disease detection. Am Heart J 1981;102:189-195.
16. Rozanski A, Diamond GA, Jones R, Forrester JS, Berman D, Morris D, Pollock BH, Freeman M, Swan HJC. A format for integrating the interpretation of exercise ejection fraction and wall motion and its application in identifying equivocal responses. J Am Coll Cardiol 1985;5:238-248.
17. Diamond GA, Forrester JS, Hirsch M, Staniloff H, VasÿR, Berman DS, Swan HJC. Application of conditional probability analysis to the clinical diagnosis of coronary artery disease. J Clin Invest 1980;65:1210-1221.
18. Santinga JT, Flora J Maple R, Brymer JF, Pitt B. The determination of the post-test likelihood for coronary artery disease using Bayes (*sic*) theorem. J Electrocardiol 1982; 15:61Ä68.
19. Dans PE, Weiner JP, Melin JA, Becker LC. Conditional probability in the diagnosis of coronary artery disease: a future tool for eliminating unnecessary testing? South Med J 76:1118-1121, 1983.
20. Christopher TD, Konstantinow G, Jones RH. Bayesian analysis of data from radionuclide angiocardiograms for diagnosis of coronary artery disease. Circulation 1984;69:65-72.
21. Detrano R, Yiannikas J, Salcedo EE, Rincon G, Go RT, Williams G, Leatherman J. Bayesian probability analysis: a prospective demonstration of its clinical utility in diagnosing coronary disease. Circulation 1984;69:541-547.
22. Lee K, Pryor D, Harrell F, Califf R, Behar V, Floyd W, Morris J, Waugh R, Whalen R, Rosati R. Predicting outcome in coronary disease. Statistical models versus expert clinicians. Am J Med 1986; 80:553-560.
23. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. NÿEngl J Med 1978;299:926-30.
24. Diamond GA, Rozanski A, Forrester JS, Morris D, Pollock BH, Staniloff HM, Berman DS, Swan HJC. A model for assessing the sensitivity and specificity of tests subject to selection bias: application to exercise radionuclide ventriculography for diagnosis of coronary artery disease. J Chron Dis 1986;39:343-55.
25. Rozanski A, Diamond GA, Berman D, Forrester JS, Morris D, Swan HJC. The declining specificity of exercise radionuclide ventriculography. New Engl J Med 1983;309:518Ä22.
26. Diamond GA. An improbable criterion of normality. Circulation 1982;66:681.
27. Diamond GA. Reverend Bayes' silent majority. An alternative factor affecting sensitivity and specificity of exercise electrocardiography. Am J Cardiol 1986;57:1175-80.
28. Knottnerus JA. The effects of disease verification and referral on the relationship between symptoms and diseases. Med Decis Making 1987;7:139-48.
29. Diamond GA. The origin of specious. Am J Cardiol 1988;62:175-7.
30. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39:207-15.
31. Diamond GA. Affirmative actions: Can the discriminant accuracy of a test be determined in the face of selection bias? Med Decis Making 1991;11:48-56.
32. Diamond GA. Off Bayes: Effect of verification bias on posterior probabilities calculated using Bayes' theorem. Med Decis Making 1992;12:22-31.
33. Morise AP, Diamond GA. A comparison of the accuracy of exercise electrocardiography in men and women using biased and unbiased populations. Am Heart J 1997;102:350-56.
34. Cecil MP, Kosinski AS, Jones MT, et al. The importance of work-up (verification) biase correction in assessing the accuracy of SPECT thallium-201 testing for the diagnosis of coronary artery disease. J Clin Epidemiol 1996;7:735-42.
35. Roger VL, Pellikka PA, Bell MR, et al. Sex and test verification bias. Impact on the diagnostic value of

exercise echocardiography. Circulation. 1997;95:405-10.
36. Budoff MJ, Diamond GA, Raggi P, Arad Y, Guerci AD, Callister TQ, Berman DS. Continuous probabilistic prediction of angiographically significant coronary artery disease using electron beam computed tomography. Circulation 2002;105:1791-1796.
37. Gray R, Begg CB, Greenes RA. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. Med Decis Making 1984;4:151-64.
38. Hunink MGM, Richardson DK, Doubilet PM, Begg CB. Testing for fetal pulmonary maturity: ROC analysis involving covariates, verification bias, and combination testing. Med Decis Making 1990;10:201-11.
39. Diamond GA. Scotched on the ROCs. Med Decis Making 1991;11:198-200.
40. Diamond GA. ROC steady: a receiver operating characteristic curve that is invariant relative to selection bias. Med Decis Making 1987;7:238-43.
41. Goodman SN. Toward evidence-based medical statistics. 1: The *P* value fallacy. Ann Intern Med 1999;130:995-1004.
42. Diamond GA, Kaul S. Prior convictions. Application of Bayes' theorem to the analysis and interpretation of clinical trials. Ann Intern Med (submitted).
43. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T. The revised COSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663-94.
43. Palmas W, Denton TA, Diamond GA. Afterimages: Bayesian enhancement of scintigraphic myocardial perfusion imaging for diagnosis of coronary artery disease. Med Decis Making 1992;12:332.
44. Diamond GA, Denton TA, Matloff JM. Fee-for-benefit. A strategy to improve the quality of health care and control costs through reimbursement incentives. J Am Coll Cardiol 1993;22:343-352.
45. Diamond GA, Denton TA, Berman DS, Cohen I. Prior restraint. A Bayesian perspective on the optimization of technology utilization for diagnosis of coronary artery disease. Am J Cardiol 1995;76:82-86.
46. Diamond GA, Wetzler HP. Evidence-based prescription drug coverage. Am J Cardiol 2001;88:767.

TABLE 1. Definition of testing terms.

	Disease State	
	Present	Absent
Positive	True Positive (TN)	False Positive (FP)

Test Result

Negative	False Negative (FN)	True Negative (TN)

$$\text{True Positive Rate} = \frac{TP}{TP + FN} = \text{Sensitivity}$$

$$\text{True Negative Rate} = \frac{TN}{TN + FP} = \text{Specificity}$$

$$\text{False Positive Rate} = \frac{FP}{TN + FP} = 1 - \text{Specificity}$$

$$\text{False Negative Rate} = \frac{FN}{TP + FN} = 1 - \text{Sensitivity}$$

TABLE 3. Prevalence of coronary artery disease according to age, sex, and symptoms

FEMALES

Age	Asymptomatic	Nonanginal Discomfort	Atypical Angina	Typical Angina	
35	0.3	1	4	26	
45	1	3	13	55	
55	3	8	32	79	
65	8	19	54	91	

MALES

Age	Asymptomatic	Nonanginal Discomfort	Atypical Angina	Typical Angina	
35	2	5	22	70	
45	6	14	46	87	
55	10	22	59	92	
65	12	28	67	94	

All values are in percent